

RESEARCH ARTICLE

Tailoring circulation type classification outcomes

Felicitas Hansen  | Danijel Belušić 

Rosby Centre, Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden

Correspondence

Felicitas Hansen, Swedish Meteorological and Hydrological Institute (SMHI), 601 76 Norrköping, Sweden.
Email: felicitas.hansen@smhi.se

Funding information

Horizon 2020 Framework Programme, Grant/Award Number: 776613; Svenska Forskningsrådet Formas, Grant/Award Number: 2019-00829

Abstract

Efforts to intercompare many existing circulation type classification (CTC) methods have found no consistency in their outcomes. Therefore, when confronted with a task to classify atmospheric circulation types, it is difficult to find clear guidelines. This study explores the ways of increasing consistency between existing methods and obtaining physically meaningful and practically useful results. By applying a range of CTC methods to sea-level pressure fields over a Scandinavian domain, it is shown that CTC methods using the same similarity measure (pattern correlation (CORR) or Euclidean distance (DIST)) have higher consistency. It is further shown that CTC outcomes can be tailored towards specific user requirements by properly manipulating the input data. Using unprocessed input data in DIST-based CTC methods frequently results in classes containing physically inconsistent members because the classification procedure is obfuscated by circulation-irrelevant information in the data. Using spatially standardized data in DIST-based methods leads to considerably improved agreement with CORR-based methods and brings high physical consistency within the individual classes. However, standardizing the input data removes too much of the circulation-relevant information and results in no clear improvement in partitioning dependent variables such as precipitation. Best performance is achieved with DIST-based methods using the input data with the spatial mean removed. This simple procedure focuses the CTC methods to use only the circulation-relevant information and hence results both in physically consistent classes and in optimally performing partitioning of dependent variables. Consequently, the recommended guideline would be to use DIST-based methods with spatial-mean-removed input data as the generally most effective classification approach.

Abbreviations: ARI, adjusted rand index; CORR, pattern correlation; CTC, circulation type classification; DIST, Euclidean distance; DJF, December–February; EV, explained variation; EV-bin, 'binarized' explained variation; GPH, geopotential height; GWT, Grosswetter types; PR, precipitation; SANDRA, simulated annealing and diversified randomization; sanom, field with spatial mean subtracted; SD, standard deviation; SD-Occ, SD of occurrence; SLP, sea level pressure; SMHI-LENS, Swedish Meteorological and Hydrological Institute large ensemble; sstand, spatially standardized field; T, 2-m temperature; tanom, temporal anomaly; Tdiff, difference in 2-m temperature between two consecutive days; Tmax/Tmin, maximum/minimum 2-m temperature; TSS, total sum of squares; U850, wind speed at 850 hPa; WSS, within-type sum of squares; xtrm, extreme.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *International Journal of Climatology* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society.

KEYWORDS

atmosphere, circulation patterns, clustering, data partitioning, data standardization, similarity measures, synoptic climatology, weather types

1 | INTRODUCTION

The large-scale atmospheric circulation is one of the most important factors influencing weather and climate conditions on different timescales. Its short- and long-term changes considerably determine both mean and extreme values of surface parameters like temperature or precipitation rates. To assess the variability of atmospheric circulation states it is often useful to condense its main aspects into a reasonably small set of dynamical situations. Circulation type classification (CTC) is often used for this purpose, and a wide range of CTC methods exist to classify mostly sea level pressure (SLP) or geopotential height (GPH) fields into characteristic circulation types (Philipp *et al.*, 2010). These methods take, for example, a time series of daily spatial fields and then try to find ‘similarities’ between them to decide which fields can be grouped together in one of a desired number of classes.

Individual circulation patterns (i.e., SLP or GPH fields) are assigned to specific classes following two principles: that the members in one class are similar to each other and the class average, and that they are dissimilar to members and averages from all other classes. The measures used to quantify these similarities vary between CTC methods, and the most commonly used are the Euclidean distance (DIST; e.g., Philipp *et al.* (2007); Michaelides *et al.* (2007); Sheridan and Lee (2011)) and the pattern correlation coefficient (CORR; e.g., Lund (1963); Kirchhofer (1974); Beck (2000)). Another way involves fuzzy-logic methods (e.g., Bardossy *et al.*, 2002); however, this is less common in the CTC context in atmospheric sciences and will not be discussed here further.

High CORR (i.e., similarity in terms of CORR) is generally found between patterns that look similar to the eye of an observer. On the other hand, two fields defined as similar based on a small DIST could look quite different to the naked eye. It might be counterintuitive and sometimes even questionable if it is desired to have fields grouped together in one class that show very different large-scale or synoptic spatial patterns. To illustrate this issue, Figure 1 shows example pairs of fields which have either low DIST (indicating similarity) but low CORR (indicating no similarity) (Figure 1a,b), or high CORR (indicating similarity) but also high DIST (indicating no similarity) (Figure 1c,d). The first pair of days shows quite diverse pressure patterns that consequently result in quite diverse local precipitation situations. The second

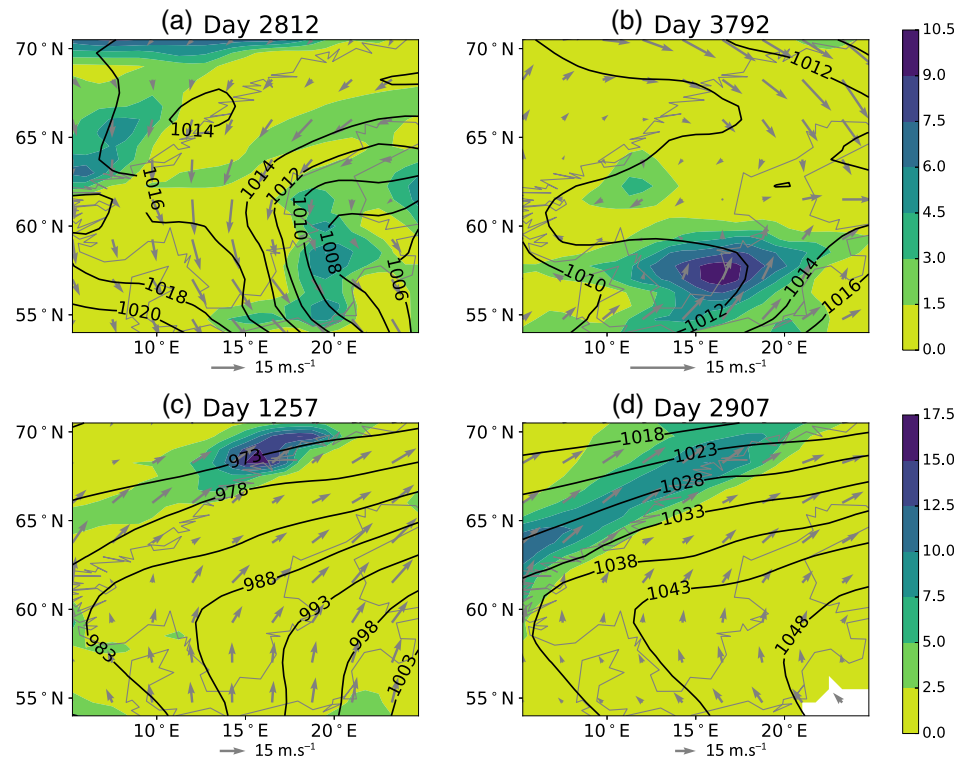
pair of days, however, shows similar pressure patterns that go along with similar precipitation fields, although the pressure magnitudes differ by several tens of hPa, thus resulting in large DIST.

It can be imagined from the examples above that the circulation types resulting from CORR-based methods can look quite different from circulation types resulting from DIST-based methods. Whether it is reasonable to group together ‘similar’ circulation patterns as in Figure 1a,b or alternatively ‘similar’ patterns as in Figure 1c,d will mostly depend on the question of interest but might sometimes be complex to evaluate.

There is a wide range of existing CTC methods and choosing the one to apply could present a challenge for a user. A comprehensive comparison of more than 20 CTC methods chosen based on their potential applicability to the European continent is provided by Philipp *et al.* (2010). Most of the methods described in Philipp *et al.* (2010)’s article are implemented in the EU COST733 Action (Philipp *et al.*, 2010) weather and circulation type catalogue which comes along with an open-source comprehensive software package that allows a user to apply the methods to their own datasets. Following Philipp *et al.* (2010), these CTC methods can roughly be divided into two groups: those that use some kind of predefined circulation types which the fields are assigned to, and those that derive circulation types as a result of the process that looks for ‘similarities’ between fields as described above. The first group includes CTC methods that provide subjectively chosen weather situations and methods where the fields are assigned to a certain circulation type based on predefined thresholds or rules. Generally, this means that there is some expert experience and knowledge of physical validity behind these methods, which is not generally the case for the methods in the second group where the partitioning of fields follows a certain algorithm. So while the advantages of automated methods are undoubted, especially in their relatively simple application to these days’ steadily growing data sets, direct interpretation of their results can be challenging. To allocate days to a certain circulation type, CTC methods that involve predefined patterns often use CORR as a similarity measure, while many of the automated methods use DIST, or even leave the choice of whether CORR or DIST is applied up to the user.

Although there is no particular CTC which could be considered ‘the truth’ (Huth *et al.*, 2007), different CTCs can be compared with each other and ranked according

FIGURE 1 First row: Example of 2 days that have low Euclidean distance for SLP, that is, are defined as ‘similar’, but also low pattern correlation ($CORR = -0.75$) for SLP which can be seen in their apparent difference and the opposite gradient direction. Second row: Example of 2 days that have large Euclidean distance for SLP, that is, are defined as ‘dissimilar’, but also high pattern correlation ($CORR = 0.96$) for SLP. Contours show SLP in hPa, colours precipitation in mm in both rows, vectors 850 hPa wind in $\text{m}\cdot\text{s}^{-1}$. Days are extracted from the climate model data described in Section 2.2 [Colour figure can be viewed at wileyonlinelibrary.com]



to various criteria. As a result of using different classification principles and similarity measures, there is low consistency between different methods in terms of their resulting circulation types. Philipp *et al.* (2010) have shown that only two out of the tested 22 methods produce similar CTC results, which, however, does not imply that these two methods perform better than the others. Furthermore, CTCs based on variables related to circulation like SLP are often intercompared with respect to their ability to partition other dependent variables like surface temperature or precipitation. Even for this intercomparison there is no clear outcome about the performance of different methods (Beck and Philipp, 2010; Casado *et al.*, 2010; Huth, 2010; Tveito, 2010).

In other fields like the data mining community it is common to use data standardization or normalization before applying a classification procedure (e.g., Keogh and Kasetty, 2003). However, the data manipulation is performed rather generally without addressing its potential physical and conceptual effects. Much more attention is given to the details of a classification method than to the data manipulation, particularly when evaluating the performance of different classification approaches. This is also common in the atmospheric and climate sciences, where the data manipulation is not frequently reported as an important step. In cases when it is reported, the most common approach is to use climatological anomalies (e.g., Meredith *et al.*, 2018; Smith and Sheridan, 2019). Some studies use different types of data standardization, which can be performed on either a spatial or a temporal

domain (e.g., Galambosi *et al.*, 1996; Ochoa-Moya *et al.*, 2020; Smith *et al.*, 2020). Here we examine different approaches for manipulating atmospheric circulation input data (SLP and GPH fields) and their effects on the CTC performance as well as their physical interpretation. We compare the effects of data manipulation to the effects of using different classification methods and show that the data manipulation alone can have larger effects on the classification performance than the details of a method.

The purpose of this article is to show that methods using the same similarity measures have higher consistency. Furthermore, by properly manipulating atmospheric circulation input data the consistency appears even between methods using different similarity measures. This finally implies that by choosing a particular combination of a similarity measure and input data pre-processing, the user can decide in advance about the general characteristics of the resulting circulation types.

2 | DATA AND METHODS

2.1 | Input data

2.1.1 | Climate model data

Analysis in this paper is based on climate model simulations of the recent past performed with the EC-Earth3 global climate model (Wyser *et al.*, 2020; Döscher *et al.*, 2021). An ensemble of 50 members has been run

for the period 1970–2014 as part of the Swedish Meteorological and Hydrological Institute large ensemble (SMHI-LENS; Wyser *et al.*, 2021) using perturbed initial conditions from six historical simulations with EC-Earth3.3.1-Veg. We have randomly selected one ensemble member and consider daily winter (December–February, DJF) data during the 45-year experiment period. Winter is chosen because circulation types affect the surface weather and climate more during the winter months and hence CTCs tend to show generally better performance (Beck and Philipp, 2010). The region of interest covers large parts of Scandinavia and is shown in Figure 1. Circulation types are computed based on daily means of sea level pressure (SLP) input data over this domain.

The SMHI-LENS provides a useful testbed for our study as it allows us to test the results in several climate realizations (i.e., ensemble members) in an environment of internally consistent circulation, thermodynamics and precipitation. The results outlined below have been tested for robustness to different ensemble members, seasons, regions and input variable (850 hPa geopotential height), and also to using year-round rather than single-season daily fields. We briefly refer to these tests in Summary and conclusions.

2.1.2 | Input data modification

As mentioned above, the most commonly used similarity measures in CTC methods are CORR and DIST, and they generally result in considerably different outputs. However, these measures can be connected after a simple modification of the input data. First, consider that spatially standardizing the data implies subtracting the spatial mean and dividing by the spatial standard deviation (SD) (e.g., Wilks, 2011). Second, note that CORR (r) by definition always uses standardized data and can be written as:

$$r(x, y) = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right) = \frac{1}{n} \sum_{i=1}^n x_i^* y_i^* = r(x^*, y^*) \quad (1)$$

where $\bar{x}$ and $\bar{y}$ are the respective spatial means of the two fields x and y , σ_x and σ_y are the spatial SDs, n is the number of data points and x^* and y^* denote the standardized versions of the fields. Furthermore, DIST (d) for standardized data can be expressed as:

$$d^2(x^*, y^*) = \sum_{i=1}^n (x_i^* - y_i^*)^2 = \sum_{i=1}^n (x_i^*)^2 + \sum_{i=1}^n (y_i^*)^2 - 2 \sum_{i=1}^n x_i^* y_i^* \quad (2)$$

where $\sum_{i=1}^n (x_i^*)^2 = \sum_{i=1}^n (y_i^*)^2 = n$. Combining Equations (1) and (2) we get (e.g., Berthold and Höppner, 2016):

$$r(x, y) = 1 - \frac{d^2(x^*, y^*)}{2n} \quad (3)$$

Note from Equation (1) that CORR-based methods are invariant under data standardization. Additionally, Equation (3) suggests that DIST-based methods, provided that they are applied to standardized data, could result in similar circulation types as CORR-based methods. Here we explore this possibility with the goal of finding potential agreement between what is generally considered as different methods that result in dissimilar circulation types.

An SLP field can be described using three components: spatial pattern (i.e., geometrical shape or orientation of isolines), spatial SD and spatial mean. The process of spatial standardization implies a considerable modification of the input data because it retains only the spatial pattern. To assist with the terminology and understanding of processes, we examine the SLP field also through the circulation framework, since the large-scale circulation (i.e., the wind vector) is proportional to the SLP gradient. Dividing the input data by the spatial SD limits the spread of values around the mean. The gradient of the standardized SLP field is naturally also scaled by the spatial SD, which similarly limits its magnitude. As a result, the standardized data loses the information about the overall gradient magnitude, implying that two SLP fields with originally very different gradient magnitudes (and hence likely dissimilar according to DIST) can become very similar after standardization provided their spatial patterns are similar. The spatial SD can thus be regarded as a proxy for the overall gradient magnitude. On the other hand, the spatial mean of the SLP field does not have a counterpart in the circulation framework since it is a constant and is eliminated through the differentiation when the gradient is calculated. This suggests that the SLP spatial mean could be considered as not relevant for the circulation. As a result, if the spatial mean is not removed from the input data, this circulation-irrelevant information will affect the CTC outcome.

Since the standardization removes the overall gradient magnitude, and the original data contains the spatial mean which is not an important circulation component, it is of interest to examine the effects of keeping the entire gradient information, both direction and magnitude, while removing the spatial mean. Therefore, we perform an additional CTC based on SLP data where only the SLP spatial mean is subtracted, an approach that has also been used by, for example, Smith *et al.* (2020).

Finally, we also use temporal anomalies as input to CTC methods. Anomalies from long-term daily climatologies are often applied to account for seasonal dependencies of the input variables (Beck and Philipp, 2010; Meredith *et al.*, 2018), and we show what can be expected from this kind of data manipulation compared with the other approaches as described above.

In summary, four types of SLP input data are used in the subsequent analysis:

1. Original data, containing all the information (hereafter: base).
2. Data with spatial mean removed (i.e., spatial anomalies), containing the information about the spatial pattern and gradient magnitude (hereafter: sanom).
3. Spatially standardized data, with spatial mean and SD removed, containing only the information about the spatial pattern (hereafter: sstand).
4. Temporal anomalies, with the daily climatological long-term mean removed (hereafter: tanom).

2.2 | CTC methods

From the large variety of existing CTC methods we focus on two in this study that serve to represent the group of predefined methods on one side and of automated methods on the other.

The first method of choice is the Grosswetter types (GWT) classification (Beck, 2000; Beck *et al.*, 2007). GWT is based on the idea that the main controlling factors of the mid-latitude climate can be reduced to three basic flow fields, which again can be traced back to the generalization of the subjective Central European Grosswetterlagen classification (Hess and Brezowsky, 1952; Gerstengarbe and Werner, 1999). Hence, GWT is based on expert experience, and as such ensures a physical soundness. The three predefined prototype fields include an idealized zonal flow field with values increasing from north to south, an idealized meridional flow field with values increasing from west to east, and an idealized cyclonic field. For each input field, CORR is computed with each of the three prototypes, and based on these coefficients and their combination, the field is assigned to a certain class. Since there is only a limited number of combinations, not all numbers of classes can be computed. For example, the set of 10 GWT classes consists of eight fields showing the main wind sectors (N, NE, E, SE, ...) in addition to one purely cyclonic and one purely anticyclonic class. By breaking down complex large-scale situations to a very simplified set of fields, the 10 GWT classes provide an intuitive way of CTC that can be applied to any region of interest (see, e.g., Beck *et al.* (2014); Broderick and Fealy (2015); Stryhal and

Huth (2017)). Other numbers of classes include mixed fields between the pure directional and (anti-)cyclonic fields. We use 10 classes for the majority of illustrations, while when evaluating the quality of data partitioning, we additionally consider the other numbers of classes that are possible in GWT: 8, 11, 16, 18, 19, 24 and 26 (see Beck, 2000; Beck *et al.*, 2007).

The second method of choice is the simulated annealing and diversified randomization (SANDRA) classification scheme (Philipp *et al.*, 2007). SANDRA is based on conventional k-means clustering but has been shown to overcome many of the limitations of this and other established automated clustering methods while at the same time being less numerically expensive. The process of simulated annealing allows fields to be temporarily assigned to certain classes even though this assignment might result in a temporary decrease of the overall data partitioning quality. In SANDRA, simulated annealing is repeated a large number of times under the concept of diversified randomization which means that the starting classes as well as the ordering of input fields are being randomized throughout the iterative process of checking and reassigning (Philipp *et al.*, 2007). When convergence is reached at the end of this process, the resulting set of classes is closer to the 'global optimum' than, for example, in standard k-means clustering. Good or even superior performance compared with other CTC methods has been shown for certain applications (e.g., Philipp *et al.*, 2007; Beck and Philipp, 2010; Philipp *et al.*, 2016; Jacobeit *et al.*, 2017; Vallorani *et al.*, 2018; Zhang *et al.*, 2018), and therefore unsurprisingly, the SANDRA method has become a widely used CTC alternative to previously established methods over various domains (e.g., Jacobeit *et al.*, 2009; Kuettel *et al.*, 2011; Murawski *et al.*, 2016; Meredith *et al.*, 2018). Any desired number of classes can be computed using SANDRA, but for consistency we use the same numbers of classes as in GWT.

The GWT classification method is based on CORR to prototype fields, while assignment of fields to classes in SANDRA is based on DIST. We have used more CTC methods from the COST733 software package which are either based on CORR or DIST to analyse their general (dis-)similarities. These methods will be mentioned only briefly here; for more detailed information, the reader is referred to the respective references given. The CORR-based methods investigated in this study are the Kirchhofer method (KIR; Kirchhofer (1974); Blair (1998)) and obliquely rotated Principal Component Analysis in T-Mode (PCT; Huth, 1993); the DIST-based methods that are used include self-organizing feature maps (SOM; e.g., Michaelides *et al.* (2007); Sheridan and Lee (2011)) and a randomized classification (RAC; see Philipp *et al.*, 2016). For two CTC methods, namely k-means

(KMN; Murtagh, 1985) and k-medoids (KMD; Kaufman and Rousseeuw, 1990) clustering, the user can choose whether CORR or DIST is used as the similarity measure, and for comparison we make use of both options.

We use daily winter (DJF) SLP data as input for the CTCs, either unmodified (base) or processed in one of the three ways outlined above (sanom, sstand, tanom). The input data type name is appended to the CTC method name for the final terminology used below (e.g., SANDRabase, SANDRAsanom, SANDRASstand and SANDRatanom).

2.3 | Comparison measures

2.3.1 | Direct comparison of CTC: Adjusted Rand index

We use the Rand index (Rand, 1971) adjusted by Hubert and Arabie (1985) (Adjusted Rand index; ARI) to evaluate the similarity between results from different CTC methods and from different input data types. The ARI is based on the classification catalogue, that is, the time series of class numbers. It quantifies co-occurrence of pairs of days between two different CTCs, CTC1 and CTC2. Following Steinley (2004), there are four options for classification of a pair of days: (a) Days in a pair belong to a single class in CTC1 and to a single class in CTC2, (b) days in a pair belong to a single class in CTC1 and to two different classes in CTC2, (c) days in a pair belong to two different classes in CTC1 and to a single class in CTC2, and (d) days in a pair belong to two different classes in CTC1 and to two different classes in CTC2.

The original Rand index (Rand, 1971) has been adjusted to the following form by Hubert and Arabie (1985):

$$ARI = \frac{\binom{N}{2} (a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{N}{2} - [(a+b)(a+c) + (c+d)(b+d)]} \quad (4)$$

a , b , c and d denote the respective numbers of pairs of days for the four options outlined above; N is the overall number of days. Equation (4) illustrates that the ARI uses all members instead of class averages only and that it does not compare classification results by measuring CORR or DIST, hence it does not depend on the same similarity measures as the CTC methods themselves. In a comparison of several classification comparison measures, Steinley (2004) mentions that the ARI is the most appropriate for this task. Philipp *et al.* (2016) point out

the robustness of the ARI in terms of different numbers of classes.

Starting from simulated classifications with perfect agreement ($ARI = 1$), Steinley (2004) used Monte Carlo simulations to estimate heuristics for determining the quality of agreement, where for example, two classifications with an $ARI > 0.90$ have ‘excellent’ correspondence, while the correspondence is ‘poor’ when $ARI < 0.65$. However, in our real-world application, we find that such high ARI values are not practically achieved despite seeing obvious improvements in agreement between different CTCs. At the same time, we find that the visual improvements in agreement between CTCs are accompanied by increases in ARI values. We therefore focus on relative ARI values, enabling comparison between different approaches and methods in our specific application.

2.3.2 | Partitioning of dependent variables

A wide range of metrics exist that aim at quantifying the CTC skill in partitioning other dependent variables like temperature, precipitation or wind. The metrics mostly describe how the variability of members within one class and the separability of class averages relate to each other. If the within-class variability is small and class separability large for a dependent variable, the CTC is considered to have some discriminatory power regarding this variable.

The first metrics used in this study is the explained variation (EV) which quantifies the fraction of variance of the dependent variable that can be accounted for by group membership (Buishand and Brandsma, 1997; Beck and Philipp, 2010; Casado *et al.*, 2010; Broderick and Fealy, 2015):

$$EV = 1 - \frac{WSS}{TSS} \quad (5)$$

WSS denotes the within-class sum of squares and TSS refers to the total sum of squares:

$$WSS = \sum_{j=1}^K \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_j)^2 \quad (6)$$

where y_{ji} denotes a member of class j and $\bar{y}_j$ its class average, K denotes the number of classes and n_j the respective number of members in each of them;

$$TSS = \sum_{i=1}^m (y_i - \bar{y})^2 \quad (7)$$

where m is the total number of days, y_i is a single day and $\bar{y}$ the overall mean.

We analyse EV for daily means of 2-m temperature (T), difference in 2-m temperature between two consecutive days (Tdiff), precipitation (PR) and wind speed at 850 hPa (U850). EV is computed separately at each grid point and then averaged over the whole domain. EV values range from 0 to 1, with higher values denoting higher discriminatory power regarding the considered variable.

CTC methods are often aimed at efficiently stratifying precipitation to understand the relationship between the large-scale circulation and the regional or local impact of precipitation events. Evaluating EV for precipitation focuses on one aspect of the two-folded nature of precipitation, that is, its intensity. An effective CTC in terms of precipitation should also address the other aspect, which is the occurrence or absence of precipitation. This capability can be measured by the SD of occurrence (SD-Occ; e.g., Tveito (2010); Broderick and Fealy (2015)):

$$SD-Occ = \sqrt{\frac{\sum_{j=1}^K (p_j - p)^2}{K-1}} \quad (8)$$

p_j refers to the relative frequency of wet days in one class, with K being the number of classes, and p the frequency of wet days for the full time series. To compute SD-Occ, the time series of precipitation intensities are translated into binary (0, 1) series using the threshold of $1 \text{ mm} \cdot \text{day}^{-1}$, with 1 and 0 denoting the occurrence of a wet and dry day, respectively. Higher SD-Occ values indicate a better distinction between the occurrence or absence of precipitation.

We use SD-Occ to also analyse how a CTC distinguishes between the occurrence and absence of extremes of certain variables. The time series of PR, maximum and minimum T and U850 are translated into binary sequences with 1 and 0 denoting, respectively, the occurrence and absence of an extreme of the associated variable. The extremes for the different variables are defined at each grid point as follows: extreme precipitation (xtrmPR) and extreme wind speed (xtrmU850) are exceedances of the local 95th percentiles, extreme maximum temperatures (xtrmTmax) are daily maximum temperatures above the local 90th percentile and extreme minimum temperatures (xtrmTmin) are daily minimum temperatures below the local 10th percentile.

Unlike EV, SD-Occ depends on the number of members per class (see Appendix). To account for this potential constraint, we compare the SD-Occ results to EV applied to the same 'binarized' time series as used for SD-Occ (referred to as EV-bin).

3 | RESULTS

3.1 | Circulation types from GWT and SANDRA

Figures 2 to 6 depict the averages of 10 classes that result from the GWT and SANDRA classifications using differently processed input data. The classes are sorted by the number of members, with Class #1 being the largest.

As defined by the method's fundamental ideas, the 10 GWT (Figure 2) classes represent the eight main wind directions together with a cyclonic (Class #8) and an anti-cyclonic (Class #7) field. The dominant (Class #1) large-scale pattern over the Scandinavian domain is a westerly

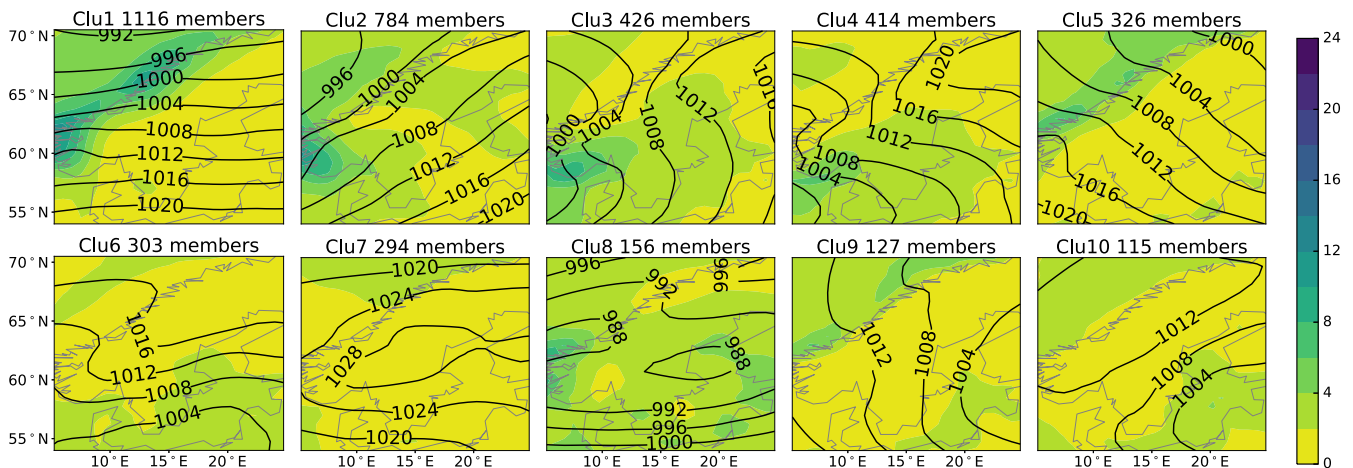


FIGURE 2 SLP (contours; interval is 4 hPa) and precipitation (colours; interval is 2 mm) for 10 class averages generated using daily DJF SLP data and the Grosswettertypes classification (GWT) [Colour figure can be viewed at wileyonlinelibrary.com]

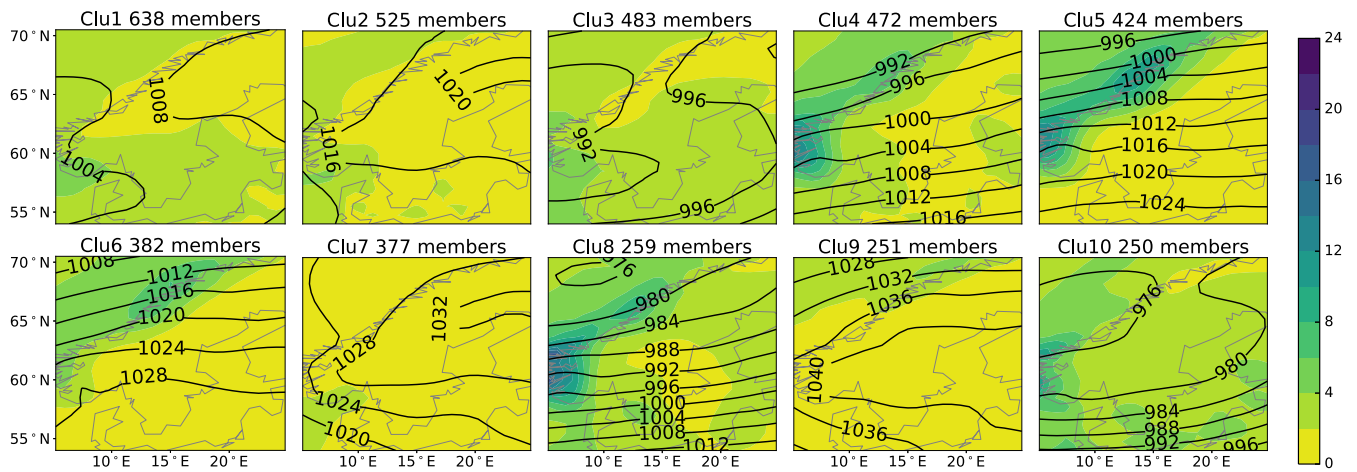


FIGURE 3 As Figure 2, except using the SANDRA classification (SANDRAbase) [Colour figure can be viewed at wileyonlinelibrary.com]

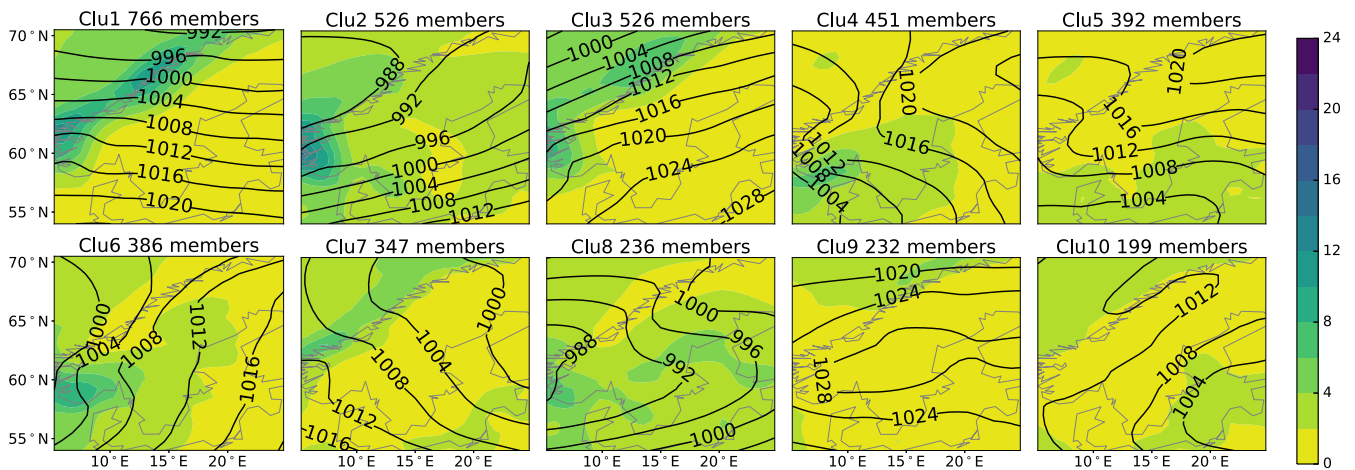


FIGURE 4 As Figure 3, except using standardized data (SANDRASstand) [Colour figure can be viewed at wileyonlinelibrary.com]

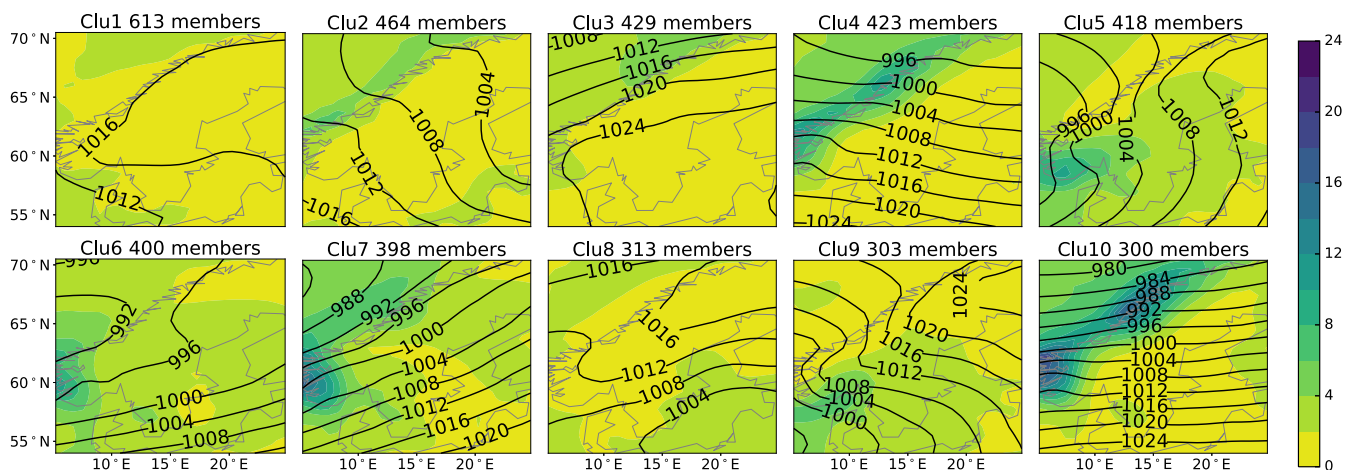


FIGURE 5 As Figure 3, except using data with the spatial mean removed (SANDRAsanom) [Colour figure can be viewed at wileyonlinelibrary.com]

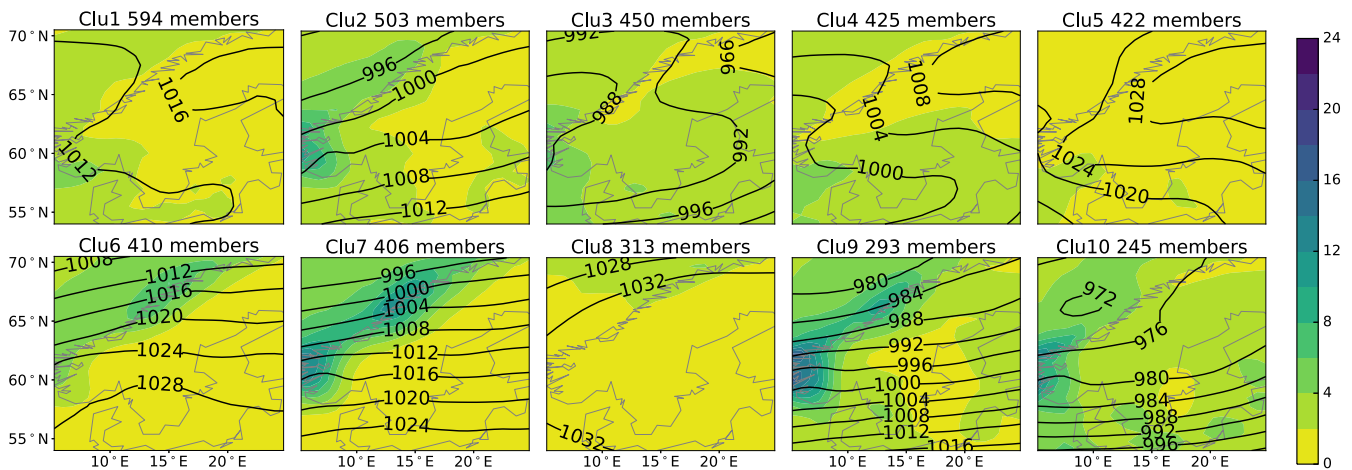


FIGURE 6 As Figure 3, except using temporal anomalies (SANDRAanom) [Colour figure can be viewed at wileyonlinelibrary.com]

flow which coincides with a precipitation maximum along the Norwegian coast. Southwesterly (Class #2), southerly (Class #3) and southeasterly (Class #4) circulation types follow in the number of days, while least days are attributed to a northeasterly flow (Class #10).

The 10 circulation types from the SANDRABase CTC build a rather different set (Figure 3). Several class averages are similar to each other, for example, showing the westerly flow in Classes #4, #5 and partly #6 and #8, although the magnitudes and pressure gradients differ remarkably between these. In addition, the patterns of Classes #1 and #3 are similar and are characterized by weak gradients, with the main difference being the SLP spatial mean. Furthermore, Classes #2, #7 and #9 are notably alike, except for a phase shift and different SLP spatial means. SANDRABase uses the original data which contains information about the spatial mean. As a result, the same spatial patterns with different spatial means can be distributed over several classes, as has just been illustrated. This equally implies that very different spatial patterns can be grouped together if they have similar spatial means, as illustrated in Figure 7. Although the SANDRABase Class #1 is characterized by a generally northeastward SLP gradient, there is a substantial number of fields assigned to this class that show the exact opposite gradient (Figure 7a). Such examples appear in other classes as well. These members are characterized by having low CORR with their class averages, which indicates low visual and hence arguably also physical correspondence (cf. Figure 1). Members that have large DIST to their class averages, however, can still have high visual correspondence (Figure 7b). Although outliers can be expected in every data set, having a considerable number of fields with opposite gradients grouped together (cf. Figure 7e) might not be considered as physically appropriate in the majority of applications.

While the SANDRABase classes are more similar among themselves than they are in general to the GWT classes, the 10 SANDRASstand class averages (Figure 4) show a very strong similarity to those of the predefined method GWT. Almost all of the circulation types from GWT have a counterpart in SANDRASstand, either corresponding even in the class numbers (Classes #1, #4, #8 and #10), or having different relative numbers of class members and hence different class numbers (e.g., GWT Class #7 corresponds to SANDRASstand Class #9). By standardizing the input SLP fields the fully automated CTC method SANDRA produces circulation types that strongly resemble the GWT circulation types. This is a noteworthy result, remembering that GWT is based on subjectively predefined types from the experience of researchers. Conceptually, the standardized data contains only the information about the spatial pattern. Consequently, the DIST measure in SANDRASstand quantifies only the distance, that is, similarity, between spatial patterns. The CORR similarity measure (used in GWT) always quantifies the similarity between spatial patterns, thus explaining the consistency between GWT and SANDRASstand. An additional consequence of SANDRASstand focusing on spatial patterns rather than spatial means is that even the most distant class members in terms of CORR and DIST show obvious correspondence to their class averages (Figure 7c, d) and that the overall physical consistency between class members is high (Figure 7e). This physical consistency is likely desired in many CTC applications and can hence be counted as an advantage of SANDRASstand compared with SANDRABase.

The class averages from SANDRASanom (Figure 5) show some resemblance to the class averages from GWT, but the similarities are not as large as between GWT and SANDRASstand. This is not surprising, recalling that the sanom data retains the information about the overall

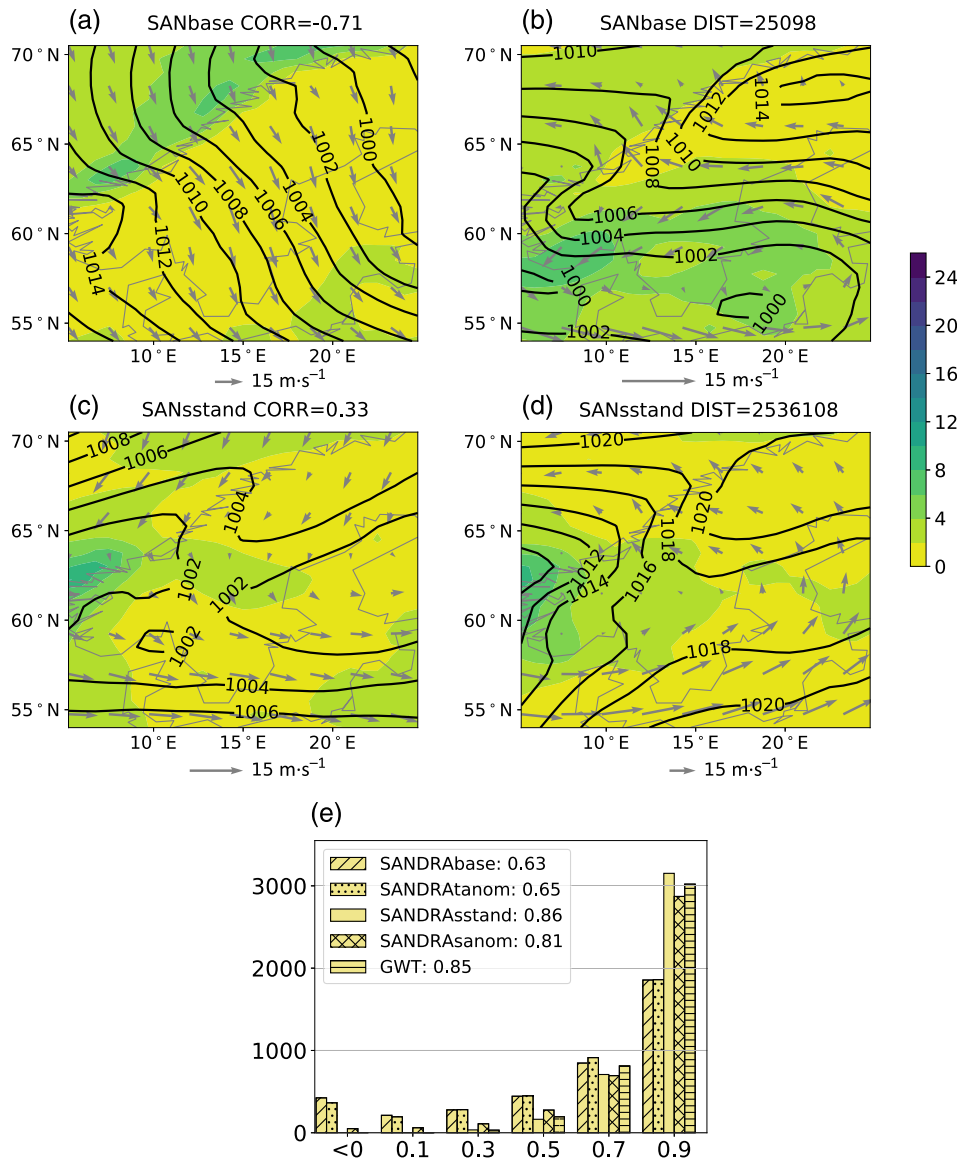


FIGURE 7 Examples of the worst-performing classes in terms of CORR in (a), (b) SANDRABase (Class #1, cf. Figure 3) and (c), (d) SANDRASstand (Class #8, cf. Figure 4). Shown are the averages over the 5% most distant members from the class averages in terms of (a), (c) CORR and (b), (d) DIST. The two classes are chosen based on the lowest average CORR of the 5% most distant members. Contours denote SLP (interval is 2 hPa), colours denote precipitation (interval is 2 mm) and vectors denote 850 hPa wind in $\text{m}\cdot\text{s}^{-1}$. (b) and (d) have a similar structure to their class averages (cf. Figure 3 and Figure 4). (e) Distribution of CORR between class members and their respective class averages for SANDRA applied to the four different data types and GWT, including all days. The values in the legend indicate the respective methods' average CORR [Colour figure can be viewed at wileyonlinelibrary.com]

gradient magnitude, which affects the final classification. From Figure 7e we see that there are considerably less members with low CORR to their respective class averages compared with SANDRABase, indicating higher physical consistency within the SANDRASanom classes. Of all the class averages shown in Figures 2 to 6, Class #10 of SANDRASanom shows the largest precipitation amounts. This suggests that SANDRASanom partitions the precipitation best, which will be further analysed in Section 3.3.

The SANDRAtanom class averages (Figure 6) have high correspondence to the class averages from SANDRABase. Like for SANDRABase, several class averages show similar patterns with varying spatial means, for example, a westerly flow in Classes #2, #6, #7 and #9. These westerly flow fields have their direct counterparts in SANDRABase Classes #4, #5, #6 and #8. The similarity between SANDRAtanom and SANDRABase also manifests in the distribution of CORR between class members

and averages (Figure 7e). A comparatively high number of members have low CORR to their class averages, which is not surprising given that the tanom data still contain information about the spatial mean. Consistent with its similarity to SANDRABase, SANDRAtanom shows generally only minor correspondence to GWT, SANDRASstand and SANDRASanom. The correspondence of SANDRAtanom and SANDRABase circulation types becomes lower for shorter time series where the amplitude of the climatological seasonal cycle that is removed is larger and when year-round instead of single-season daily data are used as CTC input (not shown).

3.2 | Correspondence to other methods

The previous section indicated a noteworthy similarity in circulation types between the GWT method, representing

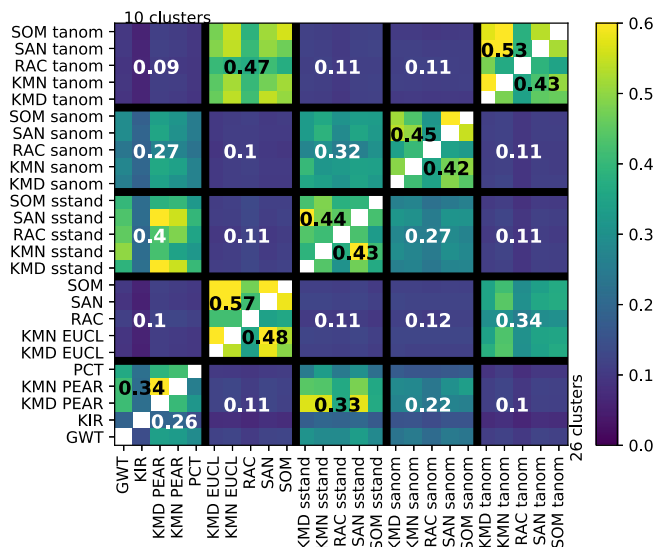


FIGURE 8 Adjusted Rand index (ARI) for CTC comparison (larger values indicate stronger similarity). Upper left half shows the ARI for 10 classes, lower right half for 26 classes. Five groups of CTCs, each including five CTC methods, are compared (from left to right in the abscissa and bottom to top in the ordinate, separated by thick black lines): (a) CORR-based methods using base data, (b) DIST-based methods using base data, (c) DIST-based methods using sstand data, (d) DIST-based methods using sanom data and (e) DIST-based methods using tanom data. The numbers indicate the averaged ARIs for the respective group comparisons [Colour figure can be viewed at wileyonlinelibrary.com]

the group of CORR-based CTC methods that use predefined circulation types, and the SANDRASstand method, representing the group of DIST-based automated methods using spatially standardized input fields. The next step is to quantify this similarity and analyse whether it also holds for other CTC methods. To accomplish that, we apply several DIST-based CTC methods to the four different types of SLP input data (base, sstand, sanom, tanom), and also several CORR-based CTC methods to the base SLP data. The results are compared between all the CTCs using the ARI and presented in a similar matrix as in Philipp *et al.* (2016) (their figure 4). Figure 8 shows the matrix, with the CTCs grouped based on the method's similarity measure (DIST or CORR) and the used type of input data. This results in five groups of CTCs: (a) CORR-based methods using base data, (b) DIST-based methods using base data, (c) DIST-based methods using sstand data, (d) DIST-based methods using sanom data and (e) DIST-based methods using tanom data. In order to show the robustness of the results, we compare the ARI for 10 and 26 classes, shown in the upper left and lower right triangle of the matrix, respectively.

As clearly seen in Figure 8, ARI is considerably higher for certain combinations of the five groups. The highest ARI is found within the groups, suggesting that the specific combination of similarity measure and data type determines the CTC result to a large extent. Further, CORR-based methods (Group 1) and DIST-based methods using sstand data (Group 3) have stronger similarity, indicating that the similarity detected between GWT and SANDRASstand in the previous section is a general characteristic of these groups. This generality is consistent with the theoretical relationship between DIST applied to sstand data and CORR (Equation (3)). When comparing to CORR-based methods (Group 1), the ARI group averages in the DIST-based methods increase by a factor of 3 when using sstand data (Group 3) rather than base data (Group 2) in the case of 26 classes, and even by a factor of 4 in the case of 10 classes.

Figure 8 further confirms the similarity, albeit somewhat weaker, between CORR-based methods (Group 1) and DIST-based methods using sanom data (Group 4), as well as between DIST-based methods using sstand (Group 3) and sanom (Group 4) data. This is in accordance with the similarity noticed in the previous section between SANDRASanom on the one hand and GWT and SANDRASstand on the other hand. Additionally notable is the correspondence between CTCs using base data (Group 2) and tanom data (Group 5). As stated in the previous section when comparing SANDRASanom to SANDRASbase, this general correspondence becomes lower for shorter time series or when using year-round data (not shown).

The KMN and KMD methods allow the user to choose CORR or DIST as the similarity measure (see Section 2.2). Comparing the circulation types that result from using CORR and DIST gives very low ARI values of around 0.1. Obtaining completely different circulation types from the same CTC method illustrates again the importance of the similarity measure in the process of assigning days to a circulation type. The ARI for KMD using CORR compared with KMD using DIST increases to 0.75 for 10 and 0.57 for 26 classes when sstand data are used.

Philipp *et al.* (2016) noted an 'unexpected low correspondence' between CTC methods in general, even between those methods that can be considered as 'closely related' like different optimization methods. Our results show that considerably higher correspondence between different methods can be obtained when sstand input data is used in DIST-based methods. The only exception which also considerably reduces the group averages of ARI is the Kirchhofer classification method (KIR). KIR consistently shows no similarity to any other method,

independent of the type of input data or the CTC similarity measure. Detecting the reasons for this exceptional behaviour is, however, beyond the scope of this paper.

3.3 | Skill in data partitioning

As mentioned before, the performance of CTCs can be examined with respect to partitioning of dependent variables. Among these variables are surface temperature and precipitation, which are often the focus of interest with CTC applications because of their potential societal or economic impact, especially during their extreme occurrences (Dixon and Mote, 2003; Yiou *et al.*, 2008; Broderick and Fealy, 2015; Kingston *et al.*, 2015; Cahynova and Huth, 2016; Li *et al.*, 2016; Vallorani *et al.*, 2018). For EV, we add to the commonly analysed variables U850 and Tdiff, the temperature difference between two consecutive days. Tdiff eliminates longer-scale temperature characteristics like its latitudinal dependency or soil and ocean thermal memory that are not immediately affected by circulation types. At the same time, Tdiff captures the same short-term variability as the circulation types based on daily data and as such is expected to better correspond to the daily changes in circulation patterns. Since a general correspondence between methods within the CTC groups was demonstrated in the previous section, we continue using SANDRA with the different types of input data and GWT as representatives of their respective groups.

Different dependent variables have different connections to the larger-scale circulation and as a result the respective values of EV and SD-Occ can vary strongly between them. However, our focus is not on the strength of these connections but rather on the relative performance of CTCs. Therefore, EV and SD-Occ values of SANDRABase are used as the reference that all other CTC methods are compared with.

The relative performance of different CTCs is quite robust for different numbers of classes (not shown, but was suggested by Figure 8). However, in order to account for potential smaller dependencies of EV and SD-Occ on the number of classes, we perform all calculations using 8, 10, 11, 16, 18, 19, 24 and 26 classes (as allowed by GWT), and present the final results as the median of these values. In summary, the final measures of EV and SD-Occ presented below follow the equation:

$$X_{rel} = \frac{\overline{X_{CTC}}}{\overline{X_{ref}}} - 1 \quad (9)$$

where X_{ref} refers to the EV or SD-Occ of SANDRABase, X_{CTC} to the EV or SD-Occ of one of the other CTCs, and the bar denotes that medians of the respective values are used. Hence, a value of $X_{rel} = 0.5$ means that the EV or SD-Occ of SANDRABase is exceeded by 50%.

Figure 9 a shows the EV of GWT, SANDRASstand, SANDRASanom and SANDRAtanom relative to SANDRABase. SANDRASanom shows the highest discriminatory power for all the variables considered here. For T and PR, SANDRASanom is the only method that performs better than SANDRABase. However, all methods lie quite close together and maximum relative EV differences reach 10% for GWT (undershooting SANDRABase for T) and SANDRASanom (exceeding SANDRABase for PR). For Tdiff and U850, three methods exceed SANDRABase by more than 20%, namely SANDRASanom, SANDRASstand and GWT, with SANDRASanom exceeding by almost 60% for Tdiff and 50% for U850. SANDRAtanom shows the most similar performance to SANDRABase for all variables, which is consistent with the ‘tanom’ group of methods showing the highest similarity with the ‘base’ group of methods in terms of the classification itself.

Remembering the physical interpretation of the different types of input data, the current results show the importance of the different physical components of the SLP field. The base (unprocessed) input data includes the entire SLP field information, including the spatial mean which is deemed generally irrelevant for circulation and as such obfuscates the CTC performance. Removing the spatial mean in sanom still retains all the circulation-related information and hence focuses the CTC procedure to only use the relevant information. The consistently better EV performance of SANDRASanom compared with SANDRABase is a clear indication of the importance of using the optimal information content in the input data. Further reduction of information such as in the typical standardization (sstand), which additionally divides by the spatial SD, removes an important aspect of the circulation, which is its overall gradient magnitude. Even though such a clustering approaches the subjective (and hence physically meaningful) classification GWT, as well as the other CORR-based CTCs, the gradient magnitude seems to play an important role. The mixed EV performance of SANDRASstand relative to SANDRABase indicates that in the tradeoff between the lack of the gradient magnitude (sstand) and the obstructive presence of the spatial mean (base) there is no clear benefit in either of the two approaches for all the dependent variables. This could explain why previous studies did not detect methods that clearly outperformed the others, because choosing a different dependent variable,

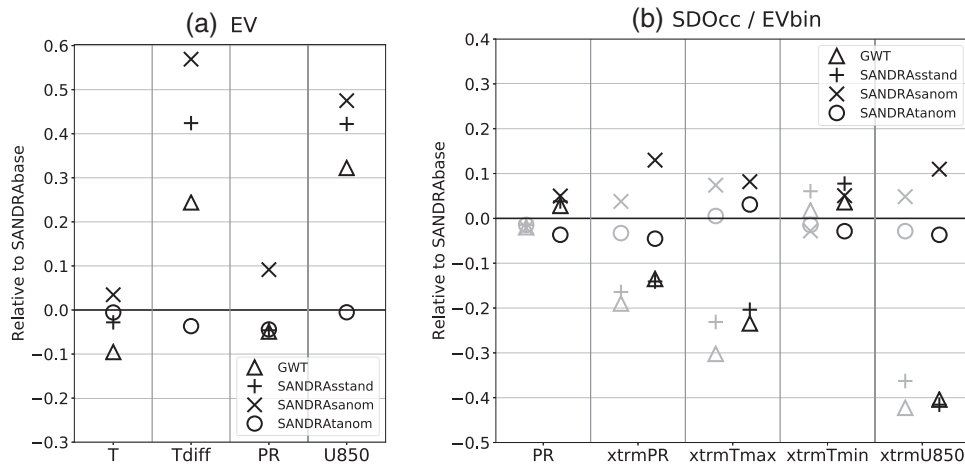


FIGURE 9 Classification performance of GWT and SANDRA using different input data types for various dependent variables assessed with three evaluation measures relative to SANDRabase. (a) Explained variation (EV) for temperature (T), temperature difference between consecutive days (Tdiff), precipitation (PR) and wind (U850); (b) SD of occurrence (SD-Occ; grey) and 'binarized' explained variation (EV-bin; black) of wet/dry day frequency (PR) and frequency of extremes in precipitation (xtrmPR), daily maximum and minimum temperature (xtrmTmax and xtrmTmin) and wind (xtrmU850)

or arguably a different region, could easily lead to a different conclusion. Consistent with their similarity, GWT performs in a similar way to SANDRASstand except that it generally has lower scores. This could be partly explained by GWT using the predefined types designed for central Europe in the northern domain, while SANDRASstand being data adaptive and hence optimizing the performance for any given data set or domain. Finally, using the temporal anomaly input data results in very small changes compared with the unprocessed data. This is because only one season is considered in this study and the influence on the CTC results are larger for year-round input data (not shown; see short discussion in Summary and Conclusions).

Figure 9 b shows SD-Occ and EV-bin of GWT, SANDRASstand, SANDRASanom and SANDRatanom relative to SANDRabase. Again it is SANDRASanom that generally shows the highest discriminatory power for the occurrence or absence of the considered variables except for xtrmTmin. SD-Occ and EV-bin agree about the order of methods for xtrmPR, xtrmTmax and xtrmU850: The GWT and SANDRASstand classifications, which emphasize the spatial pattern of the large-scale circulation, perform consistently worse than SANDRabase. Noting that it is only the SANDRASanom classification that performs better than SANDRabase, we can conclude that having combined information about the spatial pattern and gradient magnitude is key for explaining variation in PR, Tmax and U850 extremes. Having additional information about the spatial mean as included in SANDRabase, however, seems to be counterproductive for this purpose. For PR and xtrmTmin, SD-Occ and EV-bin do not show

the same level of consistency; however, the overall differences between the various methods are generally very small. Like for EV (Figure 9), the performance of SANDRatanom is generally very similar to the performance of SANDRabase, which is not surprising given the identified similarities of the sets of classes.

4 | SUMMARY AND CONCLUSIONS

Motivated by the discrepancies in results between established CTC methods, this study investigates three questions: (a) Can the discrepancies be reduced by different kinds of input data manipulation? (b) Can a specific manipulation of the atmospheric circulation input data be related to its physical effects and consequently enable a customization of CTC outcomes? (c) How do the different CTC methods and atmospheric circulation input data types partition dependent variables and is there a single recommendation for optimal CTC performance? The analysis is based on historical 45 years of daily mean winter SLP fields (i.e., atmospheric circulation patterns) over a Scandinavian domain from a member of the EC-Earth3 large ensemble.

We show that CTC methods that use the same similarity measure (DIST or CORR) in the classification procedure generally agree better with each other. By spatially standardizing the input data, DIST and CORR become related and consistency is achieved between DIST- and CORR-based methods. Therefore, the answer to the first question is yes.

To answer the second and third question, the consequences of using four different types of SLP input data were analysed: original unprocessed data, spatially standardized data, data with the spatial mean removed (which is one step of spatial standardization) and temporal anomalies from the data long-term mean. Comparison of CTC results from different combinations of methods and input data types reveals the following:

- Using unprocessed data in DIST-based methods can affect the CTC performance negatively. This is because it includes the spatial mean which is not relevant for circulation but strongly affects the minimization of DIST. It results in classes that can contain visually and physically very dissimilar members, and has mediocre skill in partitioning dependent variables. It could, however, be useful in applications where the mean SLP is important, such as for sea level studies (Sheridan *et al.*, 2019).
- Using CORR-based methods or spatially standardized data in DIST-based methods results in classes containing visually and physically highly consistent members, which is a big advantage of these approaches. Another advantage is that outcomes from different CTCs become similar to each other, as explained above. However, it does not bring general benefits in terms of partitioning performance for dependent variables. This can be explained by the standardization removing an important aspect of circulation, the overall gradient magnitude, which has different effects on different dependent variables.
- Using the spatial-mean-removed data in DIST-based methods yields the best results in terms of all aspects of CTC performance: It to a large extent keeps the physical consistency as in the CORR-based and standardized DIST-based approaches, and at the same time the explained variation of dependent variables like temperature, precipitation, wind and their respective extremes is highest. In this procedure, a CTC method uses only relevant circulation-related information about the spatial pattern and the gradient magnitude, without any extra information that could obfuscate its performance. Combining the points above, it could be loosely stated that this approach takes ‘the best of both worlds’: the physical consistency from the CORR-based and standardized DIST-based approaches, and the complete circulation information from the original DIST-based approaches that is relevant for certain dependent variables and extremes.
- Using temporal anomalies leads to only minor changes compared with the unprocessed data when daily data from one season only is used, both in terms of circulation types and partitioning performance. As discussed below, this does not apply to year-round data.

While not shown in this paper, we have tested the robustness of the results for other regions, seasons, model simulations and a different input variable, as well as for year-round instead of single-season daily data. Very similar results are obtained for all other seasons, the other input variable (850 hPa geopotential height) and different EC-Earth3 ensemble members. Using the spatial-mean-removed data outperforms other approaches in various other tested regions, including the whole of Europe, the Mediterranean and Central Canada. For a limited Central European domain, using the spatial-mean-removed data did not outperform the unprocessed data in partitioning of precipitation due to a strong signal in the mountains. Investigation of this effect is ongoing and is beyond the scope of this paper.

Using year-round instead of single-season daily data for CTC can result in a predominantly seasonal or monthly separation of circulation types. This might be desired for certain applications that involve, for example, partitioning of temperature or re-defining of the seasons (e.g., Alpert *et al.*, 2004), but undesired for others that depend on the dynamics and thermodynamics of each circulation pattern, such as partitioning of precipitation and its extremes. Our tests based on year-round data suggest that the degree to which individual SLP classes represent individual months or seasons depends largely on the amplitude and variability of the annual cycle of SLP spatial means. The unprocessed data keeps both the amplitude and variability of the annual cycle, and consequently larger differences between SLP spatial means lead to a better distinction of individual months. Temporal anomalies lack the amplitude but keep the variability of the annual cycle and as a result winter and summer are generally separated into individual clusters but individual months are not distinguished. Spatial-mean-removal or standardizing the data removes the annual cycle and hence gives lower importance to seasons and higher importance to individual circulation patterns.

Like for single-season data, the spatial-mean-removal approach also outperforms other approaches in partitioning of precipitation, wind and their respective extremes when year-round daily data were used. For the partitioning of temperature and its extremes, preliminary results differ between different regions (e.g., the best performance is obtained for temporal anomaly data in the Scandinavian domain and for the unprocessed data in the Mediterranean domain). A reason for this appears to be that CTC methods to a certain extent separate year-round data according to seasons as outlined above, which may differ between regions. The possible effect of this seasonal separation on the partitioning of temperature would be consistent with the comparatively large temperature seasonal cycle. Further investigation of this is ongoing.

In summary, physical considerations and the evaluation of the partitioning skill of the different CTC approaches suggest that using the spatial-mean-removed data in DIST-based methods is an optimal solution for the majority of applications. Subtracting the spatial mean is a very simple yet effective procedure that results in physically meaningful and well-partitioned classes.

ACKNOWLEDGEMENTS

We thank Klaus Wyser at SMHI Rossby Centre for running and making available the EC-Earth3 simulation. We further thank two anonymous reviewers for their helpful comments and suggestions. EU COST Action 733 is acknowledged for producing and making available the classification software. This study received support from the EU Horizon 2020 EUCP project (grant no. 776613) and the Formas EDUCAS project (grant no. 2019-00829).

AUTHOR CONTRIBUTIONS

Felicitas Hansen: Conceptualization; data curation; formal analysis; investigation; methodology; software; validation; visualization; writing-original draft; writing-review & editing. **Danijel Belusić:** Conceptualization; funding acquisition; investigation; methodology; project administration; supervision; validation; writing-original draft; writing-review & editing.

ORCID

Felicitas Hansen  <https://orcid.org/0000-0003-4696-3059>

Danijel Belusić  <https://orcid.org/0000-0002-5665-3866>

REFERENCES

- Alpert, P., Osetinsky, I., Ziv, B. and Shafir, H. (2004) A new seasons definition based on classified daily synoptic systems: an example for the eastern Mediterranean. *International Journal of Climatology*, 24(8), 1013–1021. <https://doi.org/10.1002/joc.1037>.
- Bardossy, A., Stehlik, J. and Caspary, H.J. (2002) Automated objective classification of daily circulation patterns for precipitation and temperature downscaling based on optimized fuzzy rules. *Climate Research*, 23(1), 11–22. <https://doi.org/10.3354/cr023011>.
- Beck, C., Jacobeit, J. and Jones, P.D. (2007) Frequency and within-type variations of large-scale circulation types and their effects on low-frequency climate variability in Central Europe since 1780. *International Journal of Climatology*, 27(4), 473–491. <https://doi.org/10.1002/joc.1410>.
- Beck, C. (2000) *Zirkulationsdynamische Variabilität im Bereich Nordatlantik-Europa seit 1780*. Würzburger Geographische Arbeiten, 95, 350 pp.
- Beck, C. and Philipp, A. (2010) Evaluation and comparison of circulation type classifications for the European domain. *Physics and Chemistry of the Earth*, 35(9–12), 374–387.
- Beck, C., Weitnauer, C. and Jacobeit, J. (2014) Downscaling of monthly PM10 indices at different sites in Bavaria (Germany) based on circulation type classifications. *Atmospheric Pollution Research*, 5(4), 741–752. <https://doi.org/10.5094/APR.2014.083>.
- Berthold, M.R. and Höppner, F. (2016) Clustering Time Series Using Euclidean Distance and Pearson Correlation. arXiv: 1601.02213 [cs.LG].
- Blair, D. (1998) The Kirchhofer technique of synoptic typing revisited. *International Journal of Climatology*, 18(14), 1625–1635.
- Broderick, C. and Fealy, R. (2015) An analysis of the synoptic and climatological applicability of circulation type classifications for Ireland. *International Journal of Climatology*, 35(4), 481–505. <https://doi.org/10.1002/joc.3996>.
- Buishand, T.A. and Brandsma, T. (1997) Comparison of circulation classification schemes for predicting temperature and precipitation in The Netherlands. *International Journal of Climatology*, 17(8), 875–889. [https://doi.org/10.1002/\(SICI\)1097-0088\(19970630\)17](https://doi.org/10.1002/(SICI)1097-0088(19970630)17).
- Cahynova, M. and Huth, R. (2016) Atmospheric circulation influence on climatic trends in Europe: an analysis of circulation type classifications from the COST733 catalogue. *International Journal of Climatology*, 36, 2743–2760. <https://doi.org/10.1002/joc.4003>.
- Casado, M.J., Pastor, M.A. and Doblas-Reyes, F.J. (2010) Links between circulation types and precipitation over Spain. *Physics and Chemistry of the Earth*, 35, 437–447. <https://doi.org/10.1016/j.pce.2009>.
- Dixon, P.G. and Mote, T.L. (2003) Patterns and causes of Atlanta's urban Heat Island-initiated precipitation. *Journal of Applied Meteorology*, 42, 1273–1284. [https://doi.org/10.1175/1520-0450\(2003\)042<1273:PACOAU>2.0.CO;2](https://doi.org/10.1175/1520-0450(2003)042<1273:PACOAU>2.0.CO;2).
- Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arneth, A., Arsouze, T., Bergmann, T., Bernadello, R., Bousetta, S., Caron, L.-P., Carver, G., Castrillo, M., Catalano, F., Cvijanovic, I., Davini, P., Dekker, E., Doblas-Reyes, F.J., Docquier, D., Echevarria, P., Fladrich, U., Fuentes-Franco, R., Gröger, M., Hardenberg, J.V., Hieronymus, J., Karami, M.P., Keskinen, J.-P., Koenigk, T., Makkonen, R., Massonnet, F., Ménéguez, M., Miller, P.A., Moreno-Chamarro, E., Nieradzick, L., van Noije, T., Nolan, P., O'Donnell, D., Ollinaho, P., van den Oord, G., Ortega, P., Prims, O.T., Ramos, A., Reerink, T., Rousset, C., Ruprich-Robert, Y., Le Sager, P., Schmith, T., Schrödner, R., Serva, F., Sicardi, V., Madsen, M.S., Smith, B., Tian, T., Tourigny, E., Uotila, P., Vancoppenolle, M., Wang, S., Wärlind, D., Willén, U., Wyser, K., Yang, S., Yepes-Arbós, X. and Zhang, Q. (2021) The EC-Earth3 Earth System model for the climate model Intercomparison project 6. *Geoscientific Model Development Discussion*, 1–90, 2021. <https://doi.org/10.5194/gmd-2020-446>.
- Galambosi, A., Duckstein, L. and Bogardi, I. (1996) Evaluation and analysis of daily atmospheric circulation patterns of the 500 hPa pressure field over the southwestern USA. *Atmospheric Research*, 40, 49–76. <https://doi.org/10.1016/0169-8095>.
- Gerstengarbe, F.W. and Werner, P.C. (1999) *Katalog der Großwetterlagen Europas (1881–1998) nach Paul Hess und Helmuth Brezowsky*. Deutscher Wetterdienst: Berichte des Deutschen Wetterdienstes. <https://books.google.se/books?id=a-hzwEACAAJ>.
- Hess, P. and Brezowsky, H. (1952) *Katalog der Grosswetterlagen Europas*. Bad Kissingen: Deutscher Wetterdienst.
- Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification*, 2(1), 193–218. <https://doi.org/10.1007/BF01908075>.
- Huth, R. (1993) An example of using obliquely rotated principal components to detect circulation types over Europe. *Meteorologische Zeitschrift*, 2, 285–293.

- Huth, R. (2010) Synoptic-climatological applicability of circulation classifications from the COST733 collection: first results. *Physics and Chemistry of the Earth*, 35(9–12), 388–394.
- Huth, R., Beck, C., Philipp, A., Demuzere, M., Ustrnul, Z., Cahynova, M., Kysely, J. and Tveito, O.E. (2007) Classifications of atmospheric circulation patterns recent advances and applications. In: Gimeno, L., GarciaHerrera, R. and Trigo, R.M. (Eds.) *Trends and Directions in Climate Research*, Vol. 1146. San Lorenzo de El Escorial, Spain: *Annals of the New York Academy of Sciences*, 7th Annual Meeting of the European-Meteorology-Society (EMS)/8th European Conference on Applications of Meteorology. <https://doi.org/10.1196/annals.1446.019>.
- Jacobeit, J., Homann, M., Philipp, A. and Beck, C. (2017) Atmospheric circulation types and extreme areal precipitation in southern central Europe. *Advances in Science and Research*, 14, 71–75. <https://doi.org/10.5194/asr-14-71-2017>.
- Jacobeit, J., Rathmann, J., Philipp, A. and Jones, P.D. (2009) Central European precipitation and temperature extremes in relation to large-scale atmospheric circulation types. *Meteorologische Zeitschrift*, 18, 397–410.
- Kaufman, L. and Rousseeuw, P.J. (1990) Wiley series in probability and statistics: applied probability and statistics. In: *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: John Wiley.
- Keogh, E. and Kasetty, S. (2003) On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4), 349–371. <https://doi.org/10.1023/A:1024988512476>.
- Kingston, D.G., Stagge, J.H., Tallaksen, L.M. and Hannah, D.M. (2015) European-scale drought: understanding connections between atmospheric circulation and meteorological drought indices. *Journal of Climate*, 28, 505–516.
- Kirchhofer, W. (1974) *Circulation of European 500 Mb Patterns*. Arbeitsberichte der Schweizerischen Meteorologischen Zentralanstalt, Swiss Meteorological Institute, <https://books.google.se/books?id=3YhSQwAACAAJ>.
- Kuettel, M., Luterbacher, J. and Wanner, H. (2011) Multidecadal changes in winter circulation-climate relationship in Europe: frequency variations, within-type modifications, and long-term trends. *Climate Dynamics*, 36, 957–972.
- Li, M., Zhang, Q. and Zhang, F. (2016) Hail day frequency trends and associated atmospheric circulation patterns over China during 1960–2012. *Journal of Climate*, 29, 7027–7044.
- Lund, I.A. (1963) Map-pattern classification by statistical methods. *Journal of Applied Meteorology*, 2, 56–65. [https://doi.org/10.1175/1520-0450\(1963\)002<0056:MPCBSM>2.0.CO;2](https://doi.org/10.1175/1520-0450(1963)002<0056:MPCBSM>2.0.CO;2).
- Meredith, E.P., Rust, H.W. and Ulbrich, U. (2018) A classification algorithm for selective dynamical downscaling of precipitation extremes. *Hydrology and Earth System Sciences*, 22, 4183–4200. <https://doi.org/10.5194/hess-22-4183-2018>.
- Michaelides, S.C., Liassidou, F. and Schizas, C.N. (2007) Synoptic classification and establishment of analogues with artificial neural networks. *Pure and Applied Geophysics*, 164, 1347–1364. <https://doi.org/10.1007/s00024-007-0222-7>.
- Murawski, A., Buerger, G., Vorogushyn, S. and Merz, B. (2016) Can local climate variability be explained by weather patterns? A multi-station evaluation for the Rhine basin. *Hydrology and Earth System Sciences*, 20, 4283–4306.
- Murtagh, F. (1985) *Multidimensional Clustering Algorithms*. Physica-Verlag HD: Compstat lectures. <https://books.google.se/books?id=1RbvAAAAAAAJ>.
- Ochoa-Moya, C.A., Cala-Pérez, Y.A., Díaz-Esteban, Y., Castro, C.L., Ordoñez-Peréz, P. and Quintanar, A.I. (2020) ISSN 2073-4433 Climatological large-scale circulation patterns over the middle Americas region. *Atmosphere*, 11, 745. <https://doi.org/10.3390/atmos11070745>, <https://www.mdpi.com/2073-4433/11/7/745>.
- Philipp, A., Della-Marta, P.M., Jacobeit, J., Fereday, D.R., Jones, P. D., Moberg, A. and Wanner, H. (2007) Long-term variability of daily North Atlantic-European pressure patterns since 1850 classified by simulated annealing clustering. *Journal of Climate*, 20(16), 4065–4095. <https://doi.org/10.1175/JCLI4175.1>.
- Philipp, A., Bartholy, J., Beck, C., Erpicum, M., Esteban, P., Fettweis, X., Huth, R., James, P., Jourdain, S., Kreienkamp, F., Krennert, T., Lykoudis, S., Michalides, S.C., Pianko-Kluczynska, K., Post, P., Alvarez, D.R., Schiemann, R., Spekat, A. and Tymvios, F.S. (2010) Cost733cat-a database of weather and circulation type classifications. *Physics and Chemistry of the Earth*, 35(9–12), 360–373.
- Philipp, A., Beck, C., Huth, R. and Jacobeit, J. (2016) Development and comparison of circulation type classifications using the COST 733 dataset and software. *International Journal of Climatology*, 36, 2673–2691. <https://doi.org/10.1002/joc.3920>.
- Rand, W.D. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846–850. <https://doi.org/10.1080/01621459.1971.10482356>.
- Sheridan, S.C. and Lee, C.C. (2011) The self-organizing map in synoptic climatological research. *Progress in Physical Geography: Earth and Environment*, 35(1), 109–119. <https://doi.org/10.1177/0309133310397582>.
- Sheridan, S.C., Lee, C.C., Adams, R.E., Smith, E.T., Pirhalla, D.E. and Ransibrahmanakul, V. (2019) Temporal modeling of anomalous coastal sea level values using synoptic climatological patterns. *Journal of Geophysical Research, Oceans*, 124, 6531–6544. <https://doi.org/10.1029/2019JC015421>.
- Smith, E.T., Lee, C.C., Barnes, B.B., Adams, R.E., Pirhalla, D.E., Ransibrahmanakul, V., Hu, C. and Sheridan, S.C. (2020) A synoptic climatological analysis of the atmospheric drivers of water clarity variability in the Great Lakes. *Journal of Applied Meteorology and Climatology*, 59, 915–935. <https://doi.org/10.1175/JAMC-D-19-0156.1>.
- Smith, E.T. and Sheridan, S.C. (2019) The influence of atmospheric circulation patterns on cold air outbreaks in the eastern United States. *International Journal of Climatology*, 39, 2080–2095. <https://doi.org/10.1002/joc.5935>.
- Steinley, D. (2004) Properties of the Hubert-Arabie adjusted rand index. *Psychological Methods*, 9, 386–396. <https://doi.org/10.1037/1082-989X.9.3.386>.
- Stryhal, J. and Huth, R. (2017) Classifications of winter euro-Atlantic circulation patterns: an intercomparison of five atmospheric reanalyses. *Journal of Climate*, 30(19), 7847–7861. <https://doi.org/10.1175/JCLI-D-17-0059.1>.
- Tveito, O.E. (2010) An assessment of circulation type classifications for precipitation distribution in Norway. *Physics and Chemistry of the Earth*, 35(9–12), 395–402. <https://doi.org/10.1016/j.pce.2010>.
- Vallorani, R., Bartolini, G., Betti, G., Crisci, A., Gozzini, B., Grifoni, D., Iannucci, M., Messeri, A., Messeri, G., Morabito, M. and

- Maracchi, G. (2018) Circulation type classifications for temperature and precipitation stratification in Italy. *International Journal of Climatology*, 38(2), 915–931. <https://doi.org/10.1002/joc.5219>.
- Wilks, D.S. (2011) Statistical methods in the atmospheric sciences. 3rd edition, Academic Press, Oxford.
- Wyser, K., Koenigk, T., Fladrich, U., Fuentes-Franco, R., Karami, M. P. and Kruschke, T. (2021) The SMHI large ensemble (SMHI-LENS) with EC-Earth3. *Geoscientific Model Development Discussion*, 1–27, 2021. <https://doi.org/10.5194/gmd-2020-428>.
- Wyser, K., van Noije, T., Yang, S., von Hardenberg, J., O'Donnell, D. and Döscher, R. (2020) On the increased climate sensitivity in the EC-Earth model from CMIP5 to CMIP6. *Geoscientific Model Development*, 13(8), 3465–3474. <https://doi.org/10.5194/gmd-13-3465-2020>.
- Yiou, P., Goubanova, K., Li, Z.X. and Nogaj, M. (2008) Weather regime dependence of extreme value statistics for summer temperature and precipitation. *Nonlinear Processes in Geophysics*, 15(3), 365–378. <https://doi.org/10.5194/npg-15-365-2008>.
- Zhang, X., Chen, D. and Yao, T. (2018) Evaluation of circulation-type classifications with respect to temperature and precipitation variations in the central and eastern Tibetan plateau. *International Journal of Climatology*, 38(13), 4938–4949. <https://doi.org/10.1002/joc.5708>.

How to cite this article: Hansen, F., & Belušić, D. (2021). Tailoring circulation type classification outcomes. *International Journal of Climatology*, 1–17. <https://doi.org/10.1002/joc.7171>

APPENDIX: SD-OCC DEPENDENCE ON THE NUMBER OF MEMBERS PER CLASS

To simply illustrate why SD-Occ depends on the numbers of members per class while EV does not, we can use a time series of 1's and 0's (representing, e.g., the occurrence of wet and dry days at a single point, respectively). In a perfect classification with two classes, all the members equal to 1 would be assigned to one class, and all the members equal to 0 to the other class. For EV (Equation (5)), both WSS' will be equal to 0 in this case, hence the perfect classification will be recognized correctly as $EV = 1$. For SD-Occ (Equation (8)), the single p_j 's will be equal to 1 in the class containing all the 1's (all members are wet days, so the wet day frequency in this class is one) and 0 for the class with all the 0's (there are no wet days in this class). But p in Equation (8), the frequency of the event occurrence for the full time series, depends on the relative occurrences of 1's and 0's and will increase with the relative number of 1's in the time series. The relative occurrences of 1's and 0's manifest as the numbers of members in the two perfect classes, and can hence affect the final SD-Occ value. So when class sizes differ a lot from each other, SD-Occ becomes larger compared with a more uniform distribution of members over the classes, and even with perfect clustering, it can have quite different values. In reality, this effect will probably be small in most cases.